university of groningen

University Medical Center Groningen

University of Groningen

How Concrete Do We Get Telling Stories?

Vossen, Piek; Caselli, Tomasso; Cybulska, Agata

Published in:
Topics in Cognitive Science

DOI:
10.1111/tops.12366

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Vossen, P., Caselli, T., & Cybulska, A. (2018). How Concrete Do We Get Telling Stories? *Topics in Cognitive Science*, *10*(3), 621-640. https://doi.org/10.1111/tops.12366

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 06-05-2023

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 10 (2018) 621–640
© 2018 The Authors Topics in Cognitive Science published by Wiley Periodicals, Inc. on behalf of Cognitive Science Society.
ISSN: 1756-8765 online
DOI: 10.1111/tops.12366

This article is part of the topic "Abstract Concepts: Structure, Processing, and Modeling," Marianna Bolognesi and Gerard Steen (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

How Concrete Do We Get Telling Stories?

Piek Vossen,[a] Tommaso Caselli,[b] Agata Cybulska[c]

[a]*Computational Lexicology and Terminology Lab, Faculty of Humanities, VU Amsterdam*
[b]*Center for Language and Cognition, Rijskuniversiteit Groningen*
[c]*Oracle, Amsterdam*

Received 21 May 2017; received in revised form 27 April 2018; accepted 2 May 2018

Abstract

Will reading different stories about the same event in the world result in a similar image of the world? Will reading the same story by different people result in a similar proxy for experiencing the story? The answer to both questions is no because language is abstract by definition and relies on our episodic experience to turn a story into a more concrete mental movie. Since our episodic knowledge differs, also the mental movie will be different. Language leaves out details, and this becomes specifically clear when building machines that read texts to represent events and to establish event relations across mentions, such as co-reference, causality, subevents, scripts, timelines, and storylines. There is a lot of information and knowledge on the event that is not in the text but is needed to reconstruct these relations and understand the story. Machines lack this knowledge and experience and likewise make explicit what it takes to understand stories from text. In this paper, we report on experiments to automatically model event descriptions and instances across different news articles. We will show that event information is scattered over the text but also varies a lot in the degree it abstracts from details, which makes establishing event identity and relations extremely difficult. The variation in granularity of event descriptions seems to vary with pragmatic communicative strategies and defines the problem at different levels of complexity.

Keywords: Event mention; Event instance; Event coreference; Storyline extraction; Computational models

Correspondence should be sent to Piek Vossen, Computational Lexicology and Terminology Lab, Faculty of Humanities, VU Amsterdam, 1081HV Amsterdam, The Netherlands. E-mail: piek.vossen@vu.nl

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The study of conceptual representations is an area rich in debates, which benefits from different fields, such as linguistics, philosophy, neuroscience, and artificial intelligence, among others. This contribution does not aim at taking a definitive stance with respect to the two major theoretical perspectives (embodied cognition hypothesis vs. the "classical" amodal hypothesis) on how concepts are represented. Instead, we address the mechanisms involved in conceptual abstraction, by focusing on how language captures and expresses them, with a particular focus on events. Through a series of computational analyses of texts, we show what it takes for computer models to reconstruct representations of events from text without having the experience and knowledge that human readers have. We explain the issues to be addressed and the possible solutions to adopt to create "intelligent" systems that can handle the identification of coreferential event mentions in- and across-documents to ultimately extract storylines, that is, temporally and logically connected sequences of events.

One assumption we make is that concepts, denoting objects or events, are abstract objects, or better stated, they are senses, that is, the constituents of Fregean propositions (Frege, 1948; Peacocke, 1992). Concepts are not the world; they model the world that consists of instances. This is even more true when considering symbolic concept systems such as natural language. The relationship between the form of a word and its meaning(s) is assumed to be arbitrary. Word forms such as "tree," "boom," "arbre," and "albero" have no direct connection to the concept, TREE, to which they are associated. Quine's inscrutability of reference (Quine, 1960) represents a philosophical argument that meaning of symbolic language is ultimately grounded in cultural and personal experience. The words "a gold digger washing sand and stones" will evoke a unique mental image in everybody's brain, none of which will exactly match the image in Fig. 1.

Communication is an exchange of personal experiences through senses associated to propositions. In addition, by adopting a pragmatic perspective, we frame our narratives by unconsciously following a set of conversational maxims, such as those formulated by Grice (1975). Grice's maxims are non-conventional (conversational) implicatures which aim at describing general principles to maximize the effective exchange of information. The (unconscious) adherence to these maxims drives our communication, allows us to leave out many details, and, most important, provides an explanation in terms of efficiency, effectiveness, and effort to communicate a message. The relation between language and the concrete perceptual world is fundamentally complex, however specific our language or our vocabulary and semantic representations may be.

Machines, or more generally, artificial agents are the perfect devices to investigate the complexity of this relationship between pragmatic aspects of communications, senses, and concepts (Searle, 1980). Machines lack personal experience and cultural background; their access to language is only through the interface of concepts. Lexical-ontological resources, such as WordNet (Fellbaum, 1998), SUMO (Niles & Pease, 2001), FrameNet (Baker, Fillmore, & Lowe, 1998), BabelNet (Navigli & Ponzetto, 2012), among others,

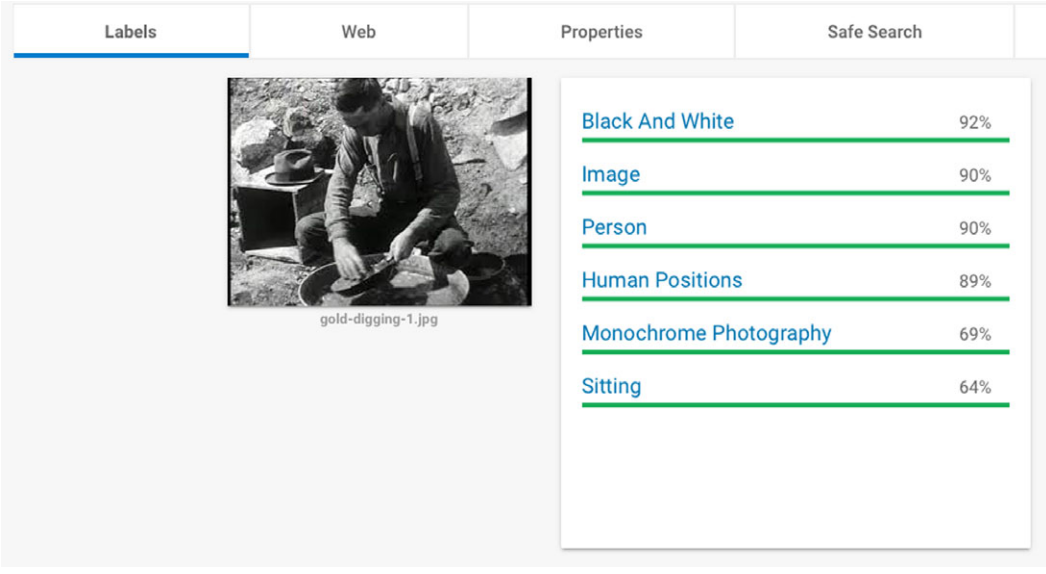

Fig. 1. Image recognition by Google vision API: https://cloud.google.com/vision/.

try to define these concepts as reflected in natural languages. But as static resources with definitions of isolated words and concepts, they lack the machinery to construe meaning in context, focusing on contextually relevant aspect and completing it with cognitive knowledge when needed. As far as perceptual knowledge is concerned, one may argue that we can now build machines that map language to image data using neural networks and large datasets. However, neural networks only create associations between visual properties of images (borders, colors, shapes, and parts) and isolated object labels. It is still a challenge for these models to derive a deeper understanding of more complex scenes and stories. The Google API will, for example, be able to tell you that the picture shown in Fig. 1 depicts a person sitting but has no understanding of the scenery: a gold digger washing sand and stones for a purpose. Images on their own do not make stories; people make stories out of any information they perceive.

We typically use language to tell stories. When people read or hear a story, they create a world in their mind that is dressedup through unique episodic perceptions and experiences not explicitly mentioned in the discourse. Language can be vague because we fill in the details through our imagination and knowledge of the world (grounded on our personal experience). Likewise, an angry man shaking his fist will look differently in everybody's personal mental "movie," but we also automatically connect his anger to other information (visually or through language), for example, on some boys and a damaged car. We try to come up with an explanation, a cause, even if that connection is not made explicitly in the text. Stories told in language abstract from perceptual experience, but also leave out many temporal, spatial, and causal relations that we tend to fill in.

Whatever the explicit message conveys, there is more not said than said. Our research question is then to find out what it takes for computers to fill these gaps and reconstruct

stories from text. How far can we get using the information that is in the text and what is needed beyond that? We approach abstraction taxonomically (Burgoon, Henderson, & Markman, 2013; Reed, 2016), in the sense that references can be made and stories can be told through very detailed construals and rich semantic language but also with the 'blink of an eye' and anything in between. Ultimately, we seek to learn, in a computational perspective, what factors determine choices for making reference at different levels of abstraction and how much can be left out for a message to still make sense. Finally, we apply our ideas to computational tasks such as detecting events in news articles, establishing event co-reference, and reconstructing storylines, that is, coherently ordered sequences of events, as a test bed.

Except for the fact that news stories report on things that happen(ed) in the world and were mostly visually and auditory perceived, they still tell us only part of the story. Typically, one needs to continue to "follow" the news and combine one document with the next to get the complete picture. Reading involves integrating information scattered across different documents over time, determining what they share, how they differ, and how information aggregates. While people have no problem doing this, computers have extreme difficulties to deal with these descriptions. These difficulties have to do with the enormous variation in the way we make reference to events, what aspects are mentioned at what granularity and specificity, and what aspects are not, but also with the fact that many details and relations are obvious and not needed for humans, based on experience and world knowledge, but are not clear and needed by machines.

We define the task of reading the news as solving several subtasks, each being non-trivial and building on top of a previous task:

1. Mentions of events in text: determine what are the relevant events mentioned in text and the components that make up event descriptions.
2. Event identity and event co-reference: establish event identity across different mentions of events.
3. Event anchoring and timeline reconstruction: anchor events in time and determine precise temporal relations between events.
4. Storyline reconstruction: select and group events that exhibit sufficient coherence and provide useful summaries with explanatory relations.

Our research on these four aspects has shown that event structures are not overtly marked in text but are the result of a construction process, which involves abstraction, information that is not present in the text but remains implicit. We claim that two main aspects are responsible for this complex construction process: the first, as already mentioned, concerns pragmatic principles of communications; and the second is related to event knowledge (Khalkhali, Wammes, & McRae, 2012; McRae & Matsuki, 2009). The first element helps us understand why some information is omitted. For instance, in case of a sentence like "two plainclothes police fatally shot the 16-year-old Kimani Gray,"[1] there is no need to mention that Kimani Gray is now dead. This information, if present, would be perceived as redundant and irrelevant. On the other hand, studies on event knowledge have shown that people use their knowledge of the world to compute

expectations for upcoming concepts, and especially events, in a discourse. There is growing experimental evidence that comprehension of sentences involves some form of anticipation for follow-up input, and that comprehension is in part driven by implicit expectations of the receiver based on his or her world knowledge.

When modeling event reconstruction from text, we follow a compositional strategy in which event structures are builtup across various mentions while aggregating components: actions, participants, locations, time, and relations between them. Our model allows us to compare event descriptions at different levels of abstraction in terms of specificity, granularity, and spatial-temporal settings. We test our model on a dataset with news articles annotated for event identity. Our attempts to map event descriptions reveal that news texts create very different stories around the same event and that it is very difficult to compare one (abstract) story with another.

In the remainder of this paper, we first discuss in section 2 the problem of event identity in relation to time: how to determine that different news articles, spread over a period of time, are making reference to the same event while abstracting from the episodic grounding in different ways (sections 2 and 3). In section 4, we discuss the problem of connecting events to form storylines on top of the extracted event structures, exhibiting the correct spatial-temporal and explanatory relations. Finally, we discuss the status of this work and conclude in section 5.

2. Event reference in language

We adopt a (neo-)Davidsonian view of events (Davidson, 2001; Higginbotham, 1995; Parsons, 2000). Events[2] are spatiotemporal entities whose participants are related to the event via thematic roles. These spatiotemporal entities are not only construed through verbal predicates but also nouns, adjectives, and prepositions can realize aspects of the event. Natural language processing adopts such a compositional vision of events: An event is a composite structure, which includes an event trigger word (i.e., a predicate) and its accompanying arguments. Traditional approaches to event detection in text start from the sentence as a unit and the predicates within the sentence: the main predicates of a clause or the heads of event-noun phrases. Given the words and phrases that can be interpreted as predicates, the next question is whether these predicates refer to the same event or not.

Event reference, or identity, is based on the kind of change, or situation, it represents; the specific participants involved; its temporal boundaries; and a spatial setting. None of these event components is by itself sufficient to establish identity: John gave Mary the book on Tuesday, John gave Mary the book on Wednesday, and Mary gave John the book on Tuesday all represent different events, although they share most or all components. Furthermore, the action itself can be described in many different ways (gets/takes/receives/borrows/buys/obtains), exhibiting different manners or perspectives. It is precisely the fact that speakers may adopt different perspectives in narrating the same episode that makes it so difficult to compare event references and establish identity.

However, people can immediately tell whether two scenes or images depict the same situation, this is extremely difficult if we summarize them in language.

Event structures require to model the accompanying arguments too: who participated, in what role, when, and where. Traditionally, this is addressed by semantic parsing and establishing the semantic role structure (Das, Chen, Martins, Schneider, & Smith, 2013). A well-known problem is, however, that not all information is given in a single sentence. Participants of events and their temporal and spatial specifics are mostly mentioned throughout the complete document. This problem becomes even more complicated if we need to compare event descriptions across texts. Different texts may exhibit different perspectives and even tell a different story for the same reality. Consider the following two fragments of text that report on the massacre of Srebrenica in the 1990s (translated from Dutch):

> On Thursday in the burning heat more than a hundred trucks and buses packed with refugees left the enclave from the Dutch UN-base Potocari. A woman and a child passed away during the trip, according to the UN. Men and boys over the age of 16 were separated from the crowd and taken away to an unknown destination. Some of them were transported to Bratunac, a city in Bosnian-Serb area to the north of the enclave. (English translation of a Dutch news article fragment, published in *Volkskrant* on 14 July 1995)

> On 11 July 1995 Serb troops under the command of General Ratko Mladi invaded the city with tanks and deported and murdered approximately 8,000 Muslim men and boys. At this time the Dutch troops known as Dutchbat were theoretically supposed to protect the enclave. Actually it was rather clear in advance that in practice it would not be possible. This event, known in the Netherlands as "the Srebrenica massacre" is seen as the worst act of genocide in Europe since the Second World War. (English translation of a fragment from the Dutch Wikipedia entry: "Het drama van Srebrenica")

The first text is written in reporting style, mentioning concrete events, more or less in their exact temporal order. The text is written shortly after the event took place (i.e., 14 July 1995). The second originates from the Dutch version of the corresponding Wikipedia entry, long after the event took place (i.e., 2004). Both texts describe the same world event, but the semantics of reference is very different. They clearly illustrate two key aspects concerning narratives and events sequences in general. The first involves abstraction. The Wikipedia text summarizes the overall event using words such as deported and murdered, leaves out details such as the trucks, the woman and the child, and adds information, interpretation, and judgment, for example, deported instead of left, trip, taken away, and transported. Such a process is a central aspect of abstraction (Burgoon et al., 2013). Furthermore, by analyzing 78 documents on Srebrenica, written either shortly after the event or with more time distance, it appears that the degree of abstraction of reference to entities, time, and location correlates with the distance in time, as shown in Fig. 2

(Cybulska & Vossen, 2010). Text written shortly after the event tends to make reference to shorter time units, individual people, smaller areas, and more local events. Text with more historical distance tends to abstract from these details, refer to longer periods, to groups rather than to individuals, to bigger areas and high-level events with more subjective interpretations, both in terms of judgment as intention.

The second aspect relates to narratives in general. Narratives have the peculiar properties of offering either social information to guide immediate decisions or general principles to make better decisions in the future (Boyd, 2009). Abstracting from the specific details of an episode is a strategy to identify regularities in events in the world. This allows our minds to loosely match events from the past to predicaments of the present to find explanations, analogies, parallels, emotional reactions to events, and their consequences.

Besides the tendency to abstract from concrete details, as well as event mentions, to more general patterns in time, there is also lot of implicit information (i.e., "not said") that must be filled in by the reader. Even for the first document, the most concrete one, we need to imagine the trucks, buses (which colors? which models?), and refugees (how many men, women, and children?), the woman and the child (how old? what do they look like?). Although, the Wikipedia text explicitly mentions 8,000 men and boys, the news article states that some of them were transported. We are also left in doubt on the precise temporal relation between the refugee transport and the separation of the men and boys. No information is given on the duration of each action or the precise distance in time (how much time between the deportations and the killings?).

When machines read such texts, they only have access to the symbolic words and expressions in the text and their meaning representations. Machines lack event knowledge, do not have expectations based on past experience, and, normally, the only access to commonsense knowledge is through language resources, which are still missing lot of relevant information. Recent studies (Granroth-Wilding & Clark, 2016; Modi, 2016;

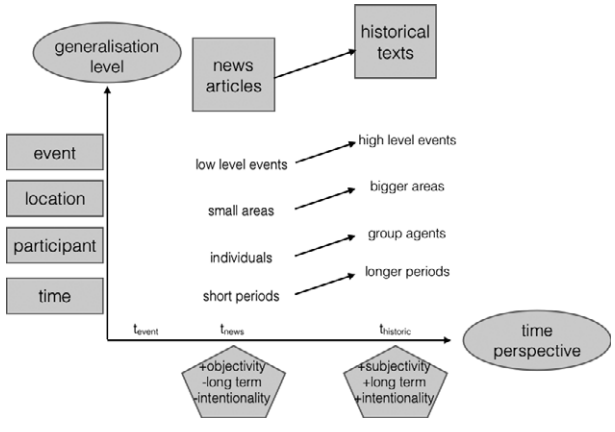

Fig. 2. Temporal perspectives on events.

Mostafazadeh et al., 2016a; Radinsky, Davidovich, & Markovitch, 2012) have shown that it is possible to develop systems that are able to predict short event sequences. Although these results are promising, such systems have been tested in limited domains (e.g., everyday activities) or by looking for specific relations between event pairs (e.g., causality). Currently, machines face the extremely challenging task to reconstruct events and stories without the material that fills the gaps.

In the next section, we explain how we automatically reconstruct events from mentions in the text, and we show how this model is used to establish event co-reference within and across documents.

3. Event coreference

To model the problem of event identity, we used an extension of the Event Co-reference Bank (ECB; Bejan & Harabagiu, 2010), where co-referential mentions of events are annotated across articles for 43 topics. The ECB+ corpus (Cybulska & Vossen, 2014) adds to the original ECB corpus new documents for each of the 43 topics. This operation is done in order to introduce extra ambiguity for each topic by selecting documents that report on similar seminal events. For example, topic 3 in ECB contained 9 news articles on an inmate (Brian Nicols) escape from a courthouse in Atlanta in 2008. For ECB+, we added 11 articles to the same topic on another inmate escape (A. J. Corneauz Jr.) from a Texas prison in 2009. The extension doubles the referential ambiguity of event mentions such as inmate escape from one to two potential world events. ECB+ contains 982 news articles with annotations for 6,833 coreferential mentions of events, mapped to 1,958 unique event instances, 4,615 human participants, 1,408 non-human ones, 1,093 time expressions, and 1,173 locations. On average, 1.8 sentences per article were annotated, predicates can potentially refer to 2.09 different events, and 3.4 different predicates refer to the same event (Ilievski, Postma, & Vossen, 2016). ECB+ is one of the richest annotated corpus for event co-reference.

As events and their components can be mentioned repeatedly in and across documents, we use a formal model, the Grounded Annotation Framework (GAF; Fokkens et al., 2013), to distinguish mentions and their instances. Each event instance is an abstraction from the specific event mentions. This allows us to lump into a single representation multiple mentions which may vary in surface realizations and the framing to narrate the event. Event instances are represented using unique identifiers, according to the Simple Event Model (SEM, Van Hage, Malaisé, Segers, Hollink, & Schreiber, 2011). Fig. 3 illustrates how mentions are mapped to unique identifiers at instance level. The challenge is to establish identity across mentions, within and across different documents. This, for example, involves normalization of relative time expressions such as Thursday to dates in the same way as 11 July 1995 represents a date, but also deciding that men and boys are the same group of entities across the two texts and that separated and taken away are related to deported. Similarly, refugees need to be matched with the crowd but also with the men, women, and children.

GAF allows us to define identity functions for each component and to combine these in a joined function.

I over two event instances e_v and e_w as the product of the identity of their components:

$$I(e_{v,w}) = \alpha I(a_{i,j}) \times \beta I(p_{t,s}) \times \gamma I(l_{m,n}) \times \delta I(t_{k,l}), \quad (1)$$

where $a_{i,j}$ represents two instances of change or situations, $p_{t,s}$ two sets of entities in a role, $l_{m,n}$ location instances, and $t_{k,l}$ the time points/periods associated to events v and w. Identity of events is then defined as a factor of the identity of each component. The constants α, β, γ, and δ allow for calibrating the contribution of each component empirically. Eq. 1 models identity as a scalar notion across partial matches such that descriptions of events can differ gradually. Identity is thus a matter of degree, where descriptions can vary in abstraction by zooming in or zooming out on details. We expect the components to contribute differently depending on the type of event; for example, for events such as change in ownership location may be less important than for disasters. As the components and details for events are spread over the complete text, we approximate event identity in three steps:

1. Establish the identity of the event components mentioned in the complete text.
2. Aggregate the component information across mentions of the same event in an instance representation of the event for a single text.
3. Establish the identity across instances from different documents by comparing their representations, using the component information across all the mentions within a single document.

The result is stored as a Composite Event Structure (CES) that consists of an event instance identifier with pointers to all the mentions in the text, participant instances with pointers to their mentions, and instance representations for place and time and their mentions. Fig. 4 shows an Resource Description Framework (RDF) representation following SEM for two event instances from the *Volkskrant* and Wikipedia, respectively, in relation to their mentions, their participants, the location, and the date.

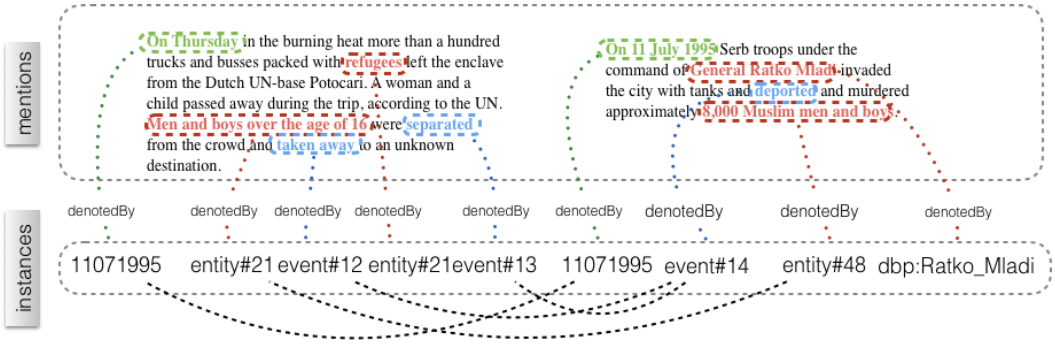

Fig. 3. Grounded Annotation Framework (GAF) representation of event mentions and instances.

After creating CESes, we determine whether they refer to the same event: where we can apply a strict matching of the components or use a similarity function. In this case, the predicates, taken_away and deported, the specific phrases for the men and boys and the locations Potocari versus Srebrenica only match through some similarity function and not literally. The dates match directly, assuming that Thursday has been correctly normalized.

We experimented with various similarity functions over the components (Cybulska & Vossen, 2015; Vossen & Cybulska, 2016) to see how they contribute to establish identity across event descriptions, in particular:

1. Similarity functions over the predicates making reference to actions
2. Granularity of participants, locations, and temporal expressions
3. Contribution of different components

In all experiments on the ECB+ dataset, we used a lemma baseline in which we consider all annotated mentions (so-called true mentions) of the same lemma as coreferential. Given the fact that predicates have a limited average ambiguity of 2.09 and there is an average lexical variation of 3.4 predicate mentions per event instance, the lemma baseline already performs reasonably well. For ECB, the lemma baseline scores 68 Recall, 84.1 Precision, and 71.1 F-score, using BLANC (Recasens & Hovy, 2011). For ECB+, the lemma baseline scores 60 Recall, 69 Precision, and 63 F-score for BLANC.

3.1. Similarity of predicates

In the similarity approach, we try to overcome lexical variation in reference (e.g., die, death, dead, pass away), using the similarity measure defined by Leacock and Chodorow (1998) exploiting WordNet.[3] Just considering co-referential events (about 10% of the annotated mentions[4]), we observed that the lemma baseline results in 18.86 Recall, 32.22 Precision, and 23.64 F-score, using the BLANC score. With optimal settings for similarity, we obtain a Recall of 20.54, a Precision of 31.05, and an F-score of 24.72. The similarity function overcomes variation but at the cost of Precision. Furthermore, lowering the

```
event#21
        gaf:denotedBy       volkskrant#taken_away
        sem:hasTime         11071995
        sem:hasActor        men_and_boys_over_the_age_of_16
        sem:hasPlace        dbp:Potocari
event#48
        gaf:denotedBy       wikipedia#deported
        sem:hasTime         11071995
        sem:hasActor        8000_muslim_men_and_boys
        sem:hasPlace        dbp:Srebrenica
```

Fig. 4. Composite event structure with instances and their mentions.

threshold for similarity results in higher Recall, but at even higher cost for Precision and lower F-score. For further experiments, we nevertheless choose a lower threshold to optimize for recall, loosely map predicates, and to use the components and properties to add Precision as is discussed below.

3.2. Granularity

Following previous works by Hobbs (1985); Mulkar-Mehta, Hobbs, and Hovy (2011); and Keet (2008), we expect the packaging of information may differ in granularity. A news article may make reference to a *conflict between Russia and Ukraine* or may report on a *Russian soldier killing an Ukrainian naval officer*. Although strongly related, these event descriptions are not coreferential, but the latter may be a subevent of the former. To establish granularity relations between event components, we created a granularity ontology. We defined 15 classes relating to granularity levels over synsets in WordNet on the basis of the ECB+ data, as shown in Fig. 5. We manually assigned these classes to 434 hypernyms in WordNet which are further linked to 11,979 more specific synsets through hyponymy relations. In addition to lexical granularity, we also considered singular and plural forms of nominal references as a granularity class. For events, we additionally use duration distributions from the database of event durations Gusev et al. (2011). Through WordNet, a large proportion of the vocabulary is thus linked to various granularity classes. Next, we used a Decision Tree in combination with the granularity features to decide on the match of event components. Our experiments showed that granularity features add 4–5% Precision: 56 Recall, 74 Precision, and 60 F-score for BLANC on ECB+, again at the price of a drop of Recall when compared to the lemma baseline.

3.3. Event components

According to Eq. 1, each component can contribute to the identity of the event, but it is unknown how much they contribute. Thus, we also experimented with including these

Event Slot	Granularity Class	Description	Synset Example
Human Participant	*gran_person*	individuals	spokesperson_1
	gran_group	groups or organizations	people_2
Location	*gran_street*	areas up to the size of a building	government_building_1
	gran_city	city districts and cities	city_district_1
	gran_country	size of a country	Upper_Egypt_1
	gran_continent	size of multiple countries	East_Africa_1
Time	*gran_second*	duration up to a minute	sec_1
	gran_min	from a minute to an hour	quarter_4
	gran_hr	from an hour up to 24 hours	hours_2
	gran_day	one to few days, less than a week	day_of_the_week_1
	gran_week	one to few weeks, less than a month	calendar_week_1
	gran_month	indication on the month level	Gregorian_calendar_month_1
	gran_season	few months	season_1
	gran_year	one or multiple years	year_1
	gran_thousands_years	thousands of years	Bronze_Age_1

Fig. 5. Granularity ontology for ECB+ event components.

components in a Decision Tree, as well in combinations with the predicates of the event mentions (Cybulska & Vossen, 2013). The predicates were matched using the WordNet similarity as described before, with a BLANC Recall of 68.1 and BLANC Precision of 71.8 for the ECB dataset. We use the WordNet similarity results for comparison as we want to maximize the recall to see the impact of the components on precision. As we have seen for the Granularity matching on ECB+, we also expect the components to add Precision.

For each component (participants, locations, and time), we used their lemmas to see if there is an exact match. For the participants, we also experimented with WordNet similarity to capture variation. We observed that all components increase Precision: 6.3 points for time (78.1 Precision), 5.5 for location (77.3 Precision), and 7.2 for participants (79 Precision). WordNet similarity for participants gives the best results for precision (79.7 Precision, 7.9 points increase) without significant loss of Recall compared to the other components and compared to WordNet similarity with just the predicates. These experiments show that components matter but that there is also a variation that needs to be captured. Time and place mentions were now matched literally but other similarity functions could be defined, like normalizing time references with respect to the document time, and defining meronymy matches for time references and for locations (e.g., Srebrenica and Bosnia).

Overall, we observed that the task itself is too artifical when compared to real-world situations in news streams, as the ECB+ is still too restricted with respect to referential ambiguity and variation, despite our efforts to increase this. We believe that our graded notion of event description is more relevant to real-life situations than the current experiments suggest. Using our formula, we can present event descriptions at different levels of abstraction and derive graded and partial identity. More research and more realistic experimental set-ups are needed to further explore this.

So far, we discussed the problem of deriving event schemas from news stories, deduplicating, and aggregating information across different mentions. In addition, we also need to capture other relations that play a role in creating larger event structures, such as subevent and causal relations. In the next section, we discuss how we further structure the event representations as coherent stories.

4. Reconstructing storylines

Events occur in context and, most important, as part of a story. This story can be told by a single document or, more commonly, is told over time by many articles, each providing bits and pieces of information. Integrating and interpreting event descriptions in a coherent and meaningful way across multiple articles can thus be framed as the task of reconstructing the corresponding story.

The storyline extraction task has three subgoals: (i) connect event descriptions in time, that is, anchoring and ordering events; (ii) identify explanatory, that is, loose cause–effect, relations between event descriptions; (iii) select relevant and salient event descriptions.

4.1. Timelines

Most approaches to structure streams of information create topic threads that develop over time, based on the sharing of participants and location. The result is a timeline, that is, a basic temporal ordering of events (Hu, Huang, & Zhu, 2014; Huang & Huang, 2013; Laparra et al., 2015b; Shahaf et al., 2013). Timeline reconstruction is not trivial. It requires resolving temporal expressions, tense, and aspect interpretation of the event descriptions, establishing the document creation time, and, finally, combining all this information to come to an interpretation of time. Even the temporal anchoring of an event, that is, establishing the precise moment an event took place, is a puzzle that requires resolving temporal information for the complete document. System performance against benchmark corpora on single document timelines, like the TempEval-3 dataset UzZaman et al. (2013), is very low, with the best system scoring 30.98 F1 Bethard (2013), and reaching only 41.41 F1 for temporal anchoring relations from raw text. In most cases, there is no, or only little, information on the temporal boundaries or values of events in the text. News texts typically do not provide precise temporal aspects and relations, leaving it to the reader to fill in the details.

The SemEval 2015 Task 4 TimeLine: Cross-document event ordering (Minard et al., 2015) introduced benchmark datasets for multi-document timeline extraction. System results vary from 7.12 to 14.31 F1 (Caselli, Fokkens, Morante, & Vossen, 2015a; Laparra, Aldabe, & Rigau, 2015a). These results highlight the complexity of the task as it combines event co-reference with temporal processing. We conducted an in-depth error analysis to investigate which modules are more prone to errors and how error propagation impacts the performance (Caselli et al., 2015b). We observed that most errors result from the event representation itself, such as participants missing from the event information because they are mentioned outside the sentence, or wrongly detected by systems. Only a minor part of the errors is due to a failure to identify the event mentions (predicates) in the documents. Extracting cross-document timelines requires systems to perform well on reconstructing complete event representations from the full text.

4.2. Storylines

Timelines do not require any further structuring of the event description information. They can be seen as long lists of event descriptions which occur at different moments in times, but there is no way of telling that the events are connected in a coherent and meaningful way. Recent attempts to connect events via meaningful coherence relations have resulted in corpora and systems for explicit causal relations (Dunietz, Levin, & Carbonell, 2015; Mirza & Tonelli, 2016; Mostafazadeh, Grealish, Chambers, Allen, & Vanderwende, 2016b). However, explicit causal relations form only a minority of the explanatory coherence relations in text. In most cases, the connecting relation needs to be added by the readers on the basis of their world knowledge. Such coherence relations are partially logically defined and partially based on experience, that is, hearing about many stories. A first attempt to learn this knowledge from large text collections resulted in the

so-called narrative schemas (Chambers & Jurafsky, 2009).The resulting structures are sets of partially ordered events (and participants) that tend to share entities but without distinguishing relevance or salience and, most important, without explanatory connection between events, except for precedence. Lacking a notion of plot structure, they result in non-coherent chains of events (Peng & Roth, 2016).

Therefore, we developed a storyline benchmark corpus: the Event Storyline Corpus (ESC) v1.0, which is a first attempt to model plot structure. Contrary to other annotation initiatives, we target newspaper articles rather than everyday activities. In this way, we can have a more realistic picture of the issues machines face to reconstruct storylines and also obtain more insight on how we, humans, tell (news) stories. ESC is a subset of the ECB+ corpus composed of 258 documents in which we formalize and annotate explanatory relations among events. The model is grounded in narratology frameworks where narratological concepts have been translated in annotation tags providing a definition and a formalization for the following components:

1. events, participants (actors), locations, and time-points (settings);
2. the anchoring of events to time and their ordering (a timeline);
3. plot/fabula relations: a set of relations between events with explanatory and predictive value(s).

ECB+ provides the basic elements of the storyline model, while ESC extends the available data by distinguishing temporal relations, marked with a so-called TLINK tag (Pustejovsky et al., 2003a), and explanatory relations, marked with the PLOT_LINK tag.[5]

PLOT_LINK annotation is conducted in two steps: First, annotators identify eligible event pairs, and then classify each relation either as a rising_action, that is, events that are circumstantial to, cause, or enable another event; or as a falling_action, that is, events denoting speculations and consequences, or the (anticipated) outcome or effect of another event. The directionality of the relations between event pairs depends on the centrality, that is, salience, of the event in the document and its positioning on an ideal concrete-abstract continuum. For example, in the following fragments, all events in bold stand in a rising_action relation with the escape event: They are concrete steps which, when summed together, may be abstracted (or generalized) from to represent a more general event, that is, an escape.

> A convicted child molester who was supposedly confined to a wheelchair overpowered two prison guards today, handcuffed them, stole their weapons and walked off wearing one of their uniforms. [. . .] Arcade Joseph Comeaux Jr. escaped just after 9 a.m.[6]

Overall 2,265 PLOT_LINK relations were annotated, with an average of 8.7 relations per document: 1,147 relations are rising_actions, while 1,118 are falling_actions. By extending the manually annotated relations with in-document event coreference, we reach 5,519 PLOT_LINKs, almost three times the average relation per document, that is, 21.39. This results in 2,653 rising_action and 2,844 falling_action relations, respectively.

Baselines systems for PLOT_LINK identification and classification show that the task is complex. So far, the best results are obtained by connecting events following the order of presentation in the text and assigning to all event pairs a rising_action relation. This resulted in 15.6 Precision, 98.8 Recall, and 26.5 F-score for the identification subtask, and 7 Precision, 94 Recall, and 14 F-score for the classification subtask. Restricting the event pairs to descriptions sharing the same temporal anchor improves Precision but at a high cost for Recall both for identification and classification (22.7 Precision and 9.7 Recall for identification, and 11.4 Precision and 5 Recall for classification, respectively). These preliminary results point out that explanatory event relations are not sequentially expressed in news texts and that it is necessary to detect these relations using external conceptual knowledge. We think that topical news archives can be used to build up this knowledge. We expect similar events to be reported using similar narrative patterns. This is not a logical necessity but a "fact of life," or our way of experiencing what happens in the world. By generalizing over individual stories, we may learn the narrative glue that underlies story coherence. We can distinguish between semantic and episodic storylines. Semantic storylines form an ontology of event schemes representing stereotypical courses of actions of topics with strong and weak causal and motivational relations. On the other hand, episodic storylines are instances that fit more or less these semantic schemes. They represent the actual course of action of a story. A final remark on the storyline model concerns its difference with respect to scripts (Schank & Abelson,1977). Storylines define probabilistic relations between events, namely explanatory or circumstatial relations in a large knowledge graph without explicitly fixing the sequences of events as in a script.

5. Conclusion

In this article, we take the position that all language utterances are abstract, that is, constituents of Fregean propositions, and that people use their episodic experience to understand and make sense of these abstract construals. We also make the argument that language also abstracts from many spatial, temporal and even causal details, which we fill in using our background knowledge. Abstractness of language makes it very difficult to judge identity across event descriptions. We illustrated this by computer models for event identity and co-reference which highlights the complex relationship between language communication and the real world. Event descriptions tend to generalize, that is, abstract, from real-world events in many different ways.

Computer models have extreme difficulties to deal with this variation. We discussed event identity, temporal anchoring of events, and storylines underlying sequences of events. The experimental results we reported must be interpreted as empirical evidence for the complexity of the different parameters involved, of the influence of time in the selection of the linguistic expressions to refer to the same event mentions, and of the impact of event knowledge and narratives in making information more or less explicit to the readers. Future work aims at learning episodic relations from large collections of news which may add the missing information for obtaining coherent event structures. To

test the progress toward this goal, we extended the ECB+ corpus with narrative relations between events. We are currently extending this annotation via crowdsourcing tasks, which allows us to have access to a larger pool of annotators. The resulting corpus will be used to evaluate our processing of the news to identify the explanatory relations between event descriptions and extract the narrative structures.

Funding

This research was carried out through funding from the VU, NWO-Spinoza, and the European Union.

NewsReader project was co-funded by the European Union as project number: 316404, FP7 Work Programme Call FP7-ICT-2011-8 Ð Objective Cooperation Research theme "Information and Communication Technologies," challenge 4.4 - Area Intelligent Information Management.

Notes

1. The sentence is extracted from the Event StoryLine Corpus v.1.0, doc 16_1.
2. In this contribution, the term "event" is used both for dynamic and static situations (Bach, 1986).
3. We extended the WordNet relations with cross-part-of-speech relations between event-nouns (e.g., payment) and verbs (pay) to extend the capacity to capture similarity, as the verb hierarchy of WordNet is shallow and not well connected.
4. It makes no sense to include non-coreferential mentions as they cannot exhibit variation.
5. For more details on the corpus, interested readers are referred to Caselli and Vossen (2017).
6. Event StoryLine Corpus v.1.0, doc 3_2.

References

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, *9*, 5–16.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley Framenet project. *Proceedings of the 17th International Conference on Computational Linguistics*-Vol. 1, (pp. 86–90). Montreal, Canada: Association for Computational Linguistics.

Bejan, C. A., & Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. In Association for Computational Linguistics (Ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1412–1422). Uppsala, Sweden: Association for Computational Linguistics.

Bethard, S. (2013), Cleartk-timeml: A minimalist approach to tempeval 2013. *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, Vol. 2, (pp. 10–14). Atlanta, GA: Association for Computational Linguistics.

Boyd, B. (2009). *On the origin of stories*. Cambridge, MA: Harvard University Press.

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science*, *8*(501), 520.

Caselli, T., Fokkens, A., Morante, R., & Vossen, P. (2015a). SPINOZA_VU: An NLP Pipeline for Cross Document Timelines. In D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 786–790). Denver, CO: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/S15-2133.

Caselli, T., & Vossen, P. (2017). The event storyline corpus: A new benchmark for causal and temporal relation extraction. In T. Caselli, B. Miller, M. van Erp, P. Vossen, M. Palmer, E. Hovy, T. Mitamura, & D. Caswell (Eds.), *Proceedings of the Events and Stories in the News Workshop* (pp. 77–86). Vancouver, Canada: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/W17.

Caselli, T., Vossen, P., vanErp, M., Fokkens, A., Ilievski, F., Bevia, R. I., Le, M., Morante, R., & Postma, M. (2015b). When it's all piling up: Investigating error propagation in an nlp pipeline. In R. Izquierdo (Ed.), *Proceedings of the Workshop on Natural Language Applications 2015*. Aache, Germany. Available at http://ceur-ws.org/Vol-1386/piling_up.pdf. Accessed Workshop on Natural Language Applications 2015.

Chambers, N., & Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: *Volume 2*, (pp. 602–610). Singapore: Association for Computational Linguistics.

Cybulska, A., & Vossen, P. (2010). Event models for historical perspectives: Determining relations between high and low level events in text based on the classification of time, location and participants. Malta.

Cybulska, A., Vossen, P. (2013). Semantic relations between events and their time, locations and participants for event coreference resolution. In G. Angelova, K. Bontcheva, & R. Mitkov (Eds.)), *Proceedings of Recent Advances in Natural Language Processing (RANLP-2013)*, no. ISSN 1313-8502. Hissar, Bulgaria: INCOMA Ltd. URL: http://aclweb.org/anthology//R/R13/R13- 1021.pdf.

Cybulska, X. X., & Vossen, (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In N. Calzolari, K. Chourki, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Odijk, & S. Piperidis (Eds.), *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)* (pp. 4545–4552). Reykjavik, Iceland: European Language Resources Association (ELRA).

Cybulska, X. X., & Vossen, (2015). Translating granularity of event slots into features for event coreference resolution. In E. Hovy, T. Mitamura, & M. Palmer (Ed.), *Proceedings of the 3rd Workshop on EVENTS: De nition, Detection, Coreference, and Representation (co-located with NAACL-2015)* (pp. 1–10). Denver, Association for Computational Linguistics.

Das, D., Chen, D., Martins, A. F. T., Schneider, N., & Smith, N. A. (2013). Frame-semantic parsing. *Computational Linguistics*, *40*(9), 56.

Davidson, D. (2001) *Essays on actions and events: Philosophical essays, volume 1*. Oxford, UK: Oxford University Press on Demand.

Dunietz, J., Levin, L., & Carbonell, J. (2015). Annotating causal language using corpus lexicography of constructions. In A. Meyers, I. Rehbein, & H. Zinsmeister (Ed.), *Proceedings of The 9th Linguistic Annotation Workshop*, (pp. 188–196). Denver, CO: Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology/W15.

Fellbaum, C. (ed.) (1998). *WordNet. An Electronic Lexical Database*. Cambridge, USA: The MIT Press.

Fokkens, A., vanErp, M., Vossen, P., Tonelli, S., vanHage, W. R., Sera ni, L., Sprugnoli, R., & Hoeksema, J. (2013). Gaf: A grounded annotation framework for events. In E. Hovy, T. Mitamura, & M. Palmer (Eds.), *Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013)*. Atlanta, GA: Association for Computational Linguistics. Available at http://aclweb.org/anthology/W/W13/W13- 1202.pdf.

Frege, G. (1948). Sense and reference. *The Philosophical Review*, *57*(209), 230.

Granroth-Wilding, M., & Clark, S. (2016). What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI-16. Proceedings of AAAI* (pp. 2727–2733). Phoenix, AZ: AAAI Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts Vol. 3 of Syntax and semantics* (pp. 41–58). New York: Academic Press.

Gusev, A., Chambers, N., Khaitan, P., Khilnani, D., Bethard, S., & Jurafsky, D. (2011). Using query patterns to learn the duration of events. In J. Bos, & S. Pulman (Eds.), *Proceedings of the Ninth International Conference on Computational Semantics (IWCS11)* (pp. 145–154). Oxford, UK: Association for Computational Linguistics.

Higginbotham, J. (1985). On semantics. *Linguistic Inquiry*, *16*(4), 547–593. Cambridge, MA: MIT Press.

Hobbs, J. R. (1985). Granularity. In A. Joshi (Eds.), *Ninth International Joint Conference on Artificial Intelligence* (pp. 432–435). San Francisco, CA: Morgan Kaufmann.

Hu, P., Huang, M.-L., & Zhu, (2014). Exploring the interactions of storylines from informative news events. *Journal of Computer Science and Technology*, *29*(502), 518.

Huang, L., & Huang, L. (2013) Optimized event storyline generation based on mixture-event-aspect model. In T. Baldwin, & A. Korhonen, (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 726–735). Seattle, WA: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/D13- 1068.

Ilievski, F., Postma, M., & Vossen, P. (2016). Semantic over tting: What 'world' do we consider when evaluating disambiguation of text? In Y. Matsumoto, & R. Prasad (Eds.), *Proceedings of 26th International Conference on Computational Linguistics* (pp. 1180–1191), COLING2016. Osaka, Japan: The COLING 2016 Organizing Committee.

Keet, C. M. (2008). A formal theory of granularity. Toward enhancing biological and applied life sciences information systems with granularity.

Khalkhali, S., Wammes, J., & McRae, K. (2012). Integrating words that refer to typical sequences of events. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*, 106.

Laparra, E., Aldabe, I., & Rigau, G. (2015a). Document level time-anchoring for timeline extraction. In C. Zong, & M. Strube, (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 358–364). Beijing, China: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/P15- 2059.

Laparra, E., Aldabe, I., & Rigau, G., (2015b). From timelines to storylines: A preliminary proposal for evaluating narratives. In T. Caselli, M. van Erp, A. Minard, M. Finlayson, B. Miller, J. Atserias, Al. Balahur & P. Vossen, (Eds.), *Proceedings of the First Workshop on Computing News Storylines* (pp. 50–55). Beijing, China: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/W15-4508.

Leacock, C., & Chodorow, M. (1998). Combining local context with wordnet similarity for word sense identification.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Lang Linguist Compass*, 3, 1417-1429. Available at http://dblp.unitrier.de/db/ journals/llc/llc3.html#McRaeM09.

Minard, A.-L., Speranza, M., Agirre, E., Aldabe, I., vanErp, M., Magnini, B., Rigau, G., Urizar, R., & Kessler, F. B. (2015). Semeval-2015 task 4: Timeline: Cross-document event ordering. In D. Cer, D. Jurgens, P. Nakov & T. Zesch (Eds.). *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 778–786) New York: The Association for Computational Linguistics.

Mirza, P., & Tonelli, S. (2016). Catena: Causal and temporal relation extraction from natural language texts. In Y. Matsumoto & R. Prasad (Eds.) *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 64–75). Osaka, Japan: The COLING 2016 Organizing Committee. Available at http://aclweb.org/anthology/C16- 1007.

Modi, A. (2016). Event embeddings for semantic script modeling. In Y. Goldberg & D. Riezler (Eds.), *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)* (pp. 75–83). Berlin, Germany: Association for Computational Linguistics.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016a). A corpus and cloze evaluation for deeper understanding of commonsense stories. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 839–849). San Diego, CA: Association for Computational Linguistics. Available at http://www. aclweb.org/anthology/ N16-1098.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., & Vanderwende, L. (2016b). Caters: Causal and temporal relation scheme for semantic annotation of event structures. In M. Palmer, T. O'Gorman, T. Mitamura & E. Hovy (Eds.) *Proceedings of the Fourth Workshop on Events* (pp. 51–61). San Diego, CA: Association for Computational Linguistics. Available: http://www.aclweb.org/anthology/ W16- 1007.

Mulkar-Mehta, R., Hobbs, J. R., & Hovy, E. (2011). Granularity in natural language discourse. In J. Bos & S. Pulman (Ed.), *Proceedings of International Conference on Computational Semantics* (pp. 360–365). Oxford, UK: Association for Computational Linguistics.

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*(217), 250.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In N. Guarino, B. Smith & C. Welty (Eds.), *Proceedings of the International Conference on Formal Ontology in Information Systems-Volume 2001* (pp. 2–9). Ogunquit, ME: ACM.

Parsons, T. (2000). Underlying states and time travel. In J. Higginbotham, F. Pianesi & A. Varzi (Eds.), *Speaking of events* (pp. 81–93). Oxford, UK: Oxford University Press.

Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: The MIT Press.

Peng, H., & Roth, D. (2016). Two discourse driven language models for semantics. In K. Erk & N. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 290–300). Berlin, Germany: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/P16- 1028.

Pustejovsky, J., Castao, J., Ingria, R., Saurì, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003a). TimeML: Robust Specification of Event and Temporal Expressions in Text. In H. Bunt, Y. Girard, E. Krahmer, R. Morante, R. Muskens, I. van der Sluís, & El. Thijsse (Eds.) *Fifth International Workshop on Computational Semantics (IWCS-5)* Tilburg, the Netherlands: IWCS-5 organizing committee.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Radinsky, K., Davidovich, S., & Markovitch, S. (2012). Learning causality for news events prediction. In A. Mille, F. Gandon & J. Misselis (Eds.) *Proceedings of the 21st International Conference on World Wide Web* (pp. 909–918). New York:ACM.

Recasens, M., & Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, *17*, (4), 485–510.

Reed, S. K. (2016). A taxonomic analysis of abstraction. *Perspectives on Psychological Science*, *11*(817), 837.

Schank, R., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*(417), 457.

Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., & Leskovec, J. (2013). Information cartography: creating zoomable, large-scale maps of information. In R. Ghani, T. Senator, P. Bradley, R. Parekh & J. He (Ed.), *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1097–1105). New York: ACM.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations.

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., & Schreiber, G. (2011). Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(128), 136.

Vossen, P., & Cybulska, A. (2016). Identity and granularity of events in text. In A. Gelbukh (Ed.), *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics conference (CICLING2016)* (pp. 501–522). Konya, Turkey: Springer's Lecture Notes in Computer Science (LNCS).